The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters

J. Hooyberghs [a,b,*], P. Van Hummelen [c], and E. Carlon [a],

[a] *Institute for Theoretical Physics, Katholieke Universiteit Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium*

[b] *Flemish Institute for Technological Research (VITO), Boeretang 200, B-2400 Mol, Belgium*

[c] *MicroArray Facility, VIB, Herestraat 49, B-3000 Leuven, Belgium*

Abstract

Quantifying interactions in DNA microarrays is of central importance for a better understanding of their functioning. Hybridization thermodynamics for nucleic acid strands in aqueous solution can be described by the so-called nearest-neighbor model, which estimates the hybridization free energy of a given sequence as a sum of dinucleotide terms. Compared with its solution counterparts, hybridization in DNA microarrays may be hindered due to the presence of a solid surface and of a high density of DNA strands. We present here a study aimed at the determination of hybridization free energies in DNA microarrays. Experiments are performed on custom Agilent slides. The solution contains a single oligonucleotide. The microarray contains spots with a perfect matching complementary sequence and other spots with one or two mismatches: in total 1006 different probe spots, each replicated 15 times per microarray. The free energy parameters are directly fitted from microarray data. The experiments demonstrate a clear correlation between hybridization free energies in the microarray and in solution. The experiments are fully consistent with the Langmuir model at low intensities, but show a clear deviation at intermediate (non-saturating) intensities. These results provide new interesting insights for the quantification of molecular

interactions in DNA microarrays.

Key words: DNA hybridization, DNA microarrays, Nearest-neighbor model

Introduction

DNA microarrays are widely used in the current research in molecular biology [1]. Such devices have several important applications [2] as for instance in gene expression profiling, in the detection of Single Nucleotide Polymorphisms, in the analysis of copy number variations and of target sequences for transcription factors. Several different platforms, either commercial or home made, are currently available. They differ by the details of fabrications (via spotting or in situ growth), the length of the sequences (oligonucleotides or long PCR fragments) and the chemistry of fixation. What all DNA microarrays have in common is the basic underlying reaction of hybridization between a nucleic acid strand in solution and a complementary strand linked covalently at a solid surface. Hybridization is characterized by a (sequence dependent) free energy difference ΔG which measures the binding affinity for the two strands to form a duplex.

In the past decades a large number of papers were dedicated to the investigation of static and dynamic properties of the hybridization between nucleic acid strands which are both floating in an aqueous solution (see [3] and references therein). Nearest neighbor models provide resonable approximation of ΔG for strands hybridizing in solution [4,5]. In these models ΔG is calculated as a sum of "stacking" parameters associated to dinucleotides [3]. The nearest neighbor model is known to be rather accurate at least for hybridization between complementary strands. The case of single internal mismatches [6] as well as the dependence

* Corresponding author.
Email addresses: jef.hooyberghs@vito.be (J. Hooyberghs), paul.vanhummelen@vib.be (P. Van Hummelen), enrico.carlon@fys.kuleuven.be (E. Carlon).

of ΔG on other parameters as the monovalent salt concentration [7] were also considered.

There has been some discussion in the literature about the relationship between hybridization in solution and hybridization in DNA microarrays. In early studies of gel pad microarrays [8] a linear relationship between microarray hybridization free energies ($\Delta G_{\mu\mathrm{array}}$) and the corresponding free energies in solution (ΔG_{sol}) was found. Recently [9] a similar relationship was observed on self-spotted codelink activated slides. Other studies on Affymetrix Genechips [10,11] report very weak correlation between $\Delta G_{\mu\mathrm{array}}$ and ΔG_{sol}. In some papers [12,13] however the same Affymetrix data could be fitted resonably well with a linearly rescaled ΔG_{sol}. Also some recent measurements of thermodynamic parameters using a temperature dependent surface plasmon resonance [14] seem to suggest a decreased $\Delta G_{\mu\mathrm{array}}$ compared to ΔG_{sol}. Clearly, as also some recent literature points out [6,15,16], more systematic physico-chemical studies are required for a better understanding of hybridization in DNA microarrays. A precise quantification of ΔG is important. Through a better understanding of molecular interactions between hybridizing strands it would be possible to turn microarrays into more precise tools for large scale genomic analysis. For instance one could estimate gene expression levels or detect mutations through an analysis based on thermodynamics instead of using empirical statistical methods.

This paper is dedicated to the investigation of the applicability of the nearest neighbor model to describe hybridization reactions in DNA microarrays, with a focus on sequences that contain isolated mismatches. Experimental results involving the hybridization of one sequence in solution with a large set of different sequences on a microarray will be presented. The stacking free energy parameters will be determined from the analysis of the behavior of the experimental fluorescent intensities measured from different spots of the microarray. We will be interested in the correlation between free energies resulting from these parameters and the equivalent quantities calculated from experimental stacking free energy parameters of nucleic acid melting in aqueous solution. The analysis of the experimental data clearly

Table 1

The oligos used as target in the four different hybridization expriments. The oligos were bought from Eurogentec in duplicate obtained from independent synthesis cycles.

Name	Sequence	Labeling
Target1	5' GTTTTCGAAGATTGGGTGGCACTGTTGTAA 3'	20-mer poly A + Cy3 on 3'
Target2	5' CAGGGCCTCGTTATCAATGGAGTAGGTTTC 3'	20-mer poly A + Cy3 on 3'
Target3	5' CTTTGTCGAGCTGGTATTTGGAGAACACGT 3'	20-mer poly A + Cy3 on 3'
Target4	5' GCTTCTCCTTAATGTCACGCACGATTTCCC 3'	20-mer poly A + Cy3 on 3'

reveals a good degree of correlation. However, a much better agreement with thermodynamic models is found if the thermodynamic parameters are directly fitted from the experimental microarray data. In addition to this tight agreement with theory, a regime is found where the data are cleary deviating from the Langmuir behavior.

This paper is organized as follows. Materials and Methods discusses the experimental setup, the thermodynamic model of hybridization and the fitting procedure. In Results and Discussion the experimental results are presented and a comparison between free energies fitted from the microarray data and their solution counterparts is done. The final part of the paper is dedicated to a general discussion in which some open issues are highlighted.

Materials and Methods

The design of the experiment

For the present study several hybridization experiments were performed, each with a single oligonucleotide sequence (referred to as the target in this paper) in solution at different

Table 2

Design of probeset: probe sequences covalently linked at the microarray surface contained up to two mismatches following the scheme shown in this table. In total there are 1006 different probe sequences, replicated 15 times in the custom 8×15K custom Agilent slide.

Nr of probes	type of mismatch	location of mismatch
1	perfect match	—
60	single mismatch (all 3 permutations)	site 6 to 25
945	double mismatch (all 9 permutations)	site 6 to 25, separated by min. 5 sites

concentrations. Four different targets were used in the experiments, and their sequences are given in Table 1. The sequences contain a 30-mer hybridizing stretch followed by a 20-mer poly(A) spacer and a Cy3 label at the 3' end of the sequence. Each target oligo was bought in duplicate, in order to check the quality of the target synthesis. In the rest of the paper we will refer to the two duplicated oligos as a and b.

The sequences printed at the microarray surfaces and referred to here as the probes, were chosen to contain up to two mismatches, following the scheme shown in Table 2. Mismatches were inserted from nucleotides 6 to 25 along the 30-mer sequences in order to avoid terminal regions. In the probes with two mismatches these were separated by at least 5 nucleotides. Given the nucleotide of the target strand there are three different possible mismatching nucleotides and 20 available positions, hence in total 60 single mismatch sequences. A similar counting for double mismatches yields 945 different sequences (see Table 2). The total number of probe sequences, including the perfect matching one, is 1006.

For each experiment one target and one 8x15K custom Agilent slide was used. This slide consists of eight identical microarrays and each of these can contain up to more than 15 thousand spots. The 1006 probe sequences were spotted in the custom array 15 times: in

Table 3

The target condition per microarray: concentration, oligo synthesis a or b, fragmentation f if applied

.

Microarray	Experiment/target 1	Experiment/target 2	Experiment/target 3	Experiment/target 4
1	10000 pM, a, f	10000 pM, a, f	10000 pM, a	1000 pM, a
2	7500 pM, a, f	5000 pM, a, f	5000 pM, a	500 pM,a
3	5000 pM, a, f	1000 pM, a, f	1000 pM, a	100 pM, a
4	2500 pM, a, f	50 pM, a, f	50 pM, a	50 pM, a
5	1000 pM, a, f	10000 pM, b, f	10000 pM, b	1000 pM, b
6	500 pM, a, f	5000 pM, b, f	5000 pM, b	500 pM, b
7	100 pM, a, f	1000 pM, b, f	1000 pM, b	100 pM, b
8	50 pM, a, f	50 pM, b, f	50 pM, b	50 pM, b

12 repicates a 30-mer poly(A) was added on the 3' side (surface side), in order to asses the effect of a sequence spacer. Three replicates contained no poly(A) spacer. The eight microarrays of one slide have to be hybridized during the same experiment, but a different target solution can be used. In the experiments the target concentrations ranged from 50 to $10,000$ pM (picomolar) according to the scheme given in Table 3. In experiment 1 only target a was used, while in the experiments 2, 3 and 4 both replicated targets (a and b) were used. Finally, in experiment 1 and 2 a fragmentation of the target was performed before hybridization (see section on hybridization protocol for details).

The four 30-mer target sequences were selected from fragments of human genes having a

GC content ranging from 43% to 50%. A criterion for selecting the target sequences was the requirement that the probes constructed following the scheme in Table 2 would yield a roughly flat histogram of mismatch types, so that all mismatches are approximately equally present in the experiments.

Hybridization protocol and scanning

For the experiments we used the commercially available Agilent platform and followed a standard protocol with Agilent products, as described below. (The target oligonucleotides were OliGold® from Eurogentec, Seraing, Belgium). Hybridization mixtures contained one target oligonucleotide with a 3' Cy3 endlabeling diluted in nuclease-free water to the final concentration together with 5 μL 10x blocking agent and 25 μL 2x GEx hybridization buffer HI-RPM. Unfortunately Agilent Techologies does not disclose the precise composition of the hybridization buffer in the content of salt and other chemicals. In experiment 1 and 2 the addition of the hybridization buffer was proceeded by a fragmentation step, 1 μL fragmentation buffer was added followed by an incubation of 30 min at 60°C. This fragmentation buffer is customarily used in Agilent hybridization platforms and produces targets sequences of reduced length in order to speed up the hybridization reaction. Too long sequences, as obtained from biological extracts, e.g. from reverse transcription of mRNA samples, have a reduced hybridization efficiency due to steric hindrance. By comparing experiments with and without fragmentation we found that the fragmentation step has little effect on the results. (More information can be found in the online supplementary material.) The hybridization mixture was centrifuged at 13000 rpm for 1 min and each microarray of the 8x15K custom Agilent slides was loaded with 40 μL. The hybridization occurred in an Agilent oven at 65°C for 17 hours with rotor setting 10 and the washing was performed according to the manufacturer's instructions. The arrays were scanned on an Agilent scanner (G2565BA) at 5 μm resolution, high+low laser intensity and further processed using Agilent Feature Extraction

Software (GE1 v5 95 Feb07) which performs automatic gridding, intensity measurement, background subtraction and quality checks.

Thermodynamic Model

In the Langmuir model the dynamics of hybridization is described by a rate equation for θ, the fraction of hybridized probes from a spot as follows

$$\frac{d\theta}{dt} = ck_1(1-\theta) - k_{-1}\theta \tag{1}$$

where c is the target concentration and k_1 and k_{-1} are the attachment and detachment rates. The equilibrium value for θ can be obtained from the condition $d\theta_{eq}/dt = 0$. Using the link between the rates and equilibrium constants, i.e. $k_1/k_{-1} = e^{-\Delta G/RT}$, with ΔG the hybridization free energy, R the gas constant and T the temperature one finds

$$\theta_{eq} = \frac{c\ e^{-\Delta G/RT}}{1 + c\ e^{-\Delta G/RT}} \tag{2}$$

which is the so-called Langmuir isotherm. To link this isotherm to the measured quantities one assumes that the fraction of hybridized probes is linearly related to the measured fluorescent intensity measured from a spot, which yields

$$I = \frac{Ac\ e^{-\Delta G/RT}}{1 + c\ e^{-\Delta G/RT}} \tag{3}$$

Here I is the background-subtracted intensity, where the background subtraction, as explained above is done by Agilent Feature Extraction software. In the rest of the paper we will no longer explicitly state that the intensities are background subtracted, and will simply refer to them as intensities. A is a constant which is an overall scale factor. Far from chemical saturation, i.e. when only a small fraction of surface sequences is hybridized (i.e.

$c\, e^{-\Delta G/RT} \ll 1$) one can neglect the denominator in Eq. (2) to get:

$$I \approx Ac\, e^{-\Delta G/RT} \tag{4}$$

In the nearest neighbor model the hybridization free energy of perfect complementary strands is approximated as a sum of dinucleotide terms. For instance:

$$\Delta G \begin{pmatrix} \text{ATCCT} \\ \text{TAGGA} \end{pmatrix} = \Delta G \begin{pmatrix} \text{AT} \\ \text{TA} \end{pmatrix} + \Delta G \begin{pmatrix} \text{TC} \\ \text{AG} \end{pmatrix} + \Delta G \begin{pmatrix} \text{CC} \\ \text{GG} \end{pmatrix} + \Delta G \begin{pmatrix} \text{CT} \\ \text{GA} \end{pmatrix} + \Delta G_{\text{init}} \tag{5}$$

where ΔG_{init} is an initiation parameter. Since we will only consider differences of ΔG between a perfect matching hybridization and a hybridization with one or multiple mismatches (see Eq. (7)), this initiation parameter will not contribute and it is omitted in the rest of the paper. For DNA/DNA hybrids, symmetries reduce the number of independent parameters to 10 [3]. The nearest neighbor model can be extended to include single internal mismatches; as an example we consider the free energy of a stretch with an internal mismatch of CT type

$$\Delta G \begin{pmatrix} \text{AT}\underline{\text{C}}\text{CT} \\ \text{TA}\underline{\text{T}}\text{GA} \end{pmatrix} = \Delta G \begin{pmatrix} \text{AT} \\ \text{TA} \end{pmatrix} + \Delta G \begin{pmatrix} \text{T}\underline{\text{C}} \\ \text{A}\underline{\text{T}} \end{pmatrix} + \Delta G \begin{pmatrix} \text{G}\underline{\text{T}} \\ \text{C}\underline{\text{C}} \end{pmatrix} + \Delta G \begin{pmatrix} \text{CT} \\ \text{GA} \end{pmatrix} \tag{6}$$

The mismatching nucleotides are underlined and for notational reasons the mismatch is always put in the second part of the dinucleotide (which requires the use of symmetry like here in dinucleotide term three) . There are 12 types of mismatches and 4 types of flanking nucleotide pairs, hence in total there are 48 mismatch parameters of dinucleotide type.

There are several possible ways of extracting the $48 + 10$ dinucleotide parameters from the experimental data. One can either fit the full Langmuir isotherm (Eq. (2)) , or for experiments at sufficiently low concentrations one could consider the limiting case of Eq. (4) . In addition, the parameters could be extracted either from an experiment at fixed concentration c, by comparing the intensities of different probe sequences, or from experiments at different concentrations by analyzing the intensities of identical probe sequences over a concentration

range. As argued in the next sections the optimal strategy in our experimental setup is to fit Eq. (4) at a fixed low concentration (the supplementary online material discusses other strategies).

Equation (4) contains the constant A which is an overall scale factor relating the hybridization probability to the actual measured fluorescence intensity. This quantity may fluctuate from experiment to experiment. For instance, the optical scanning influences A, as this is proportional to the laser intensity used. Also hybridizations in different slides might occur at slightly varying conditions and there can be small differences in the manifacturing of the slides. In the rest of this paper we will focus on relative intensities and relative free energies, i.e. for each microarray we will use the perfect match of that microarray as a point of reference. We denote the logarithmic ratios of the intensities with the perfect match intensity as

$$y_i = \ln I_i - \ln I_{PM} = -\frac{\Delta G - \Delta G_{PM}}{RT} \equiv -\frac{\Delta\Delta G}{RT} \tag{7}$$

for which the exact value of A is irrelevant and we only need to consider the relative free energy differences $\Delta\Delta G$ (which is for each probe a positive number). In $\Delta\Delta G$ of a duplex, only dinucleotide parameters which are flanking a mismatch remain, the other parameters cancel out in the subtraction. E.g. from Eq. (5) and (6) one gets

$$\Delta\Delta G\begin{pmatrix}\mathrm{AT\underline{C}CT}\\ \mathrm{TA\underline{T}GA}\end{pmatrix} = \Delta G\begin{pmatrix}\mathrm{T\underline{C}}\\ \mathrm{A\underline{T}}\end{pmatrix} + \Delta G\begin{pmatrix}\mathrm{G\underline{T}}\\ \mathrm{C\underline{C}}\end{pmatrix} - \Delta G\begin{pmatrix}\mathrm{TA}\\ \mathrm{AT}\end{pmatrix} - \Delta G\begin{pmatrix}\mathrm{AC}\\ \mathrm{TG}\end{pmatrix} \tag{8}$$

In this equation the lower strand refers to the target sequence in solution, which is fixed. The upper strand is that of the probe sequence attached to the solid surface. Hence, the $\Delta\Delta G$ of a duplex contains two mismatch dinucleotide parameters and two matching dinucleotide parameters per mismatch. This holds for sequences that contain more mismatches as long as the nearest neighbor model is valid, which we assume in our setup since mismatches are

separated at least by five base pairs. The model can now be written as

$$y_i = \sum_{\alpha=1}^{58} X_{i\alpha} \frac{\Delta G_\alpha}{RT} \tag{9}$$

where α is the index running over the 58 possible dinucleotide parameters and X is a frequency matrix, whose elements $X_{i\alpha}$ are the number of times the dinucleotide parameter α enters in $\Delta\Delta G$ of probe sequence i. With a simple extension of matrices and vectors one can rewrite the problem as

$$\vec{y} = X\vec{\omega} \tag{10}$$

where we have defined $\omega_\alpha = \Delta G_\alpha/RT$. Having written the problem in Eq. (10) as a linear one, we can now apply the standard approach to find the optimal values of the parameters. The procedure consists in minimizing $S = (\vec{y} - X\vec{\omega})^2$, which amounts to solving the following linear equation

$$X^T(\vec{y} - X\vec{\omega}) = 0 \tag{11}$$

where X^T is the transpose of X.

Degeneracies of $\vec{\omega}$

To obtain $\vec{\omega}$ from Eq. (11) one has to invert the 58×58 matrix $X^T X$. In the case that $X^T X$ is not invertible one applies a singular value decomposition [18]. In the present case the matrix is not invertible. Zero eigenvalues of the matrix $X^T X$ come from reparametrizations that leave the physically accessible parameters $\Delta\Delta G$ invariant. It is known, indeed, that the dinucleotide mismatch parameters are not uniquely determined [17,18], as these parameters are entering in the expression for the total ΔG in pairs (see Eq. (6)). For instance, a

reparametrization of the type:

$$\Delta G' \begin{pmatrix} \mathrm{x}\ \underline{\mathrm{C}} \\ \mathrm{x}'\ \underline{\mathrm{T}} \end{pmatrix} = \Delta G \begin{pmatrix} \mathrm{x}\ \underline{\mathrm{C}} \\ \mathrm{x}'\ \underline{\mathrm{T}} \end{pmatrix} + \varepsilon \qquad \Delta G' \begin{pmatrix} \mathrm{y}\ \underline{\mathrm{T}} \\ \mathrm{y}'\ \underline{\mathrm{C}} \end{pmatrix} = \Delta G \begin{pmatrix} \mathrm{y}\ \underline{\mathrm{T}} \\ \mathrm{y}'\ \underline{\mathrm{C}} \end{pmatrix} - \varepsilon \tag{12}$$

for every pair of complementary nucleotides x, x' and y, y' leaves the total ΔG invariant, as it can be verified directly from Eq. (6). Similar reparametrizations are possible for mismatches of type AG, AC and TG. Next to these there are three more invariances which involve a reparametrization of both mismatch and matching dinucleotide parameters. Hence one has at least 7 zero eigenvalues in $X^T X$. A more detailed discussion of degeneracies of $X^T X$ can be found in the supplementary online material.

Results and Discussion

Control of the quality of the experiments

As a control of the reproducibility of the result we consider the intensities correlation between analogous spots in replicated experiments. The replicated hybridizations were carried out on two microarrays of the same slide, with two identical but separately synthesized and labeled target oligos, at the same manually prepared concentration in solution, see Table 3. Figure 1 is an example thereof. It shows correlation plots between two replicated hybridizations. Two plots are shown, one with the full 15K intensities (left) and one in which the median of the intensities of the 15 replicated spots are taken (right). In the former some data spreading is observed, which is greatly reduced when the median over 15 replicated spots is taken. Note that the experimental data do not align perfectly on the diagonal of the graph, this may be attributed to the manual preparation of the solutions or to differences in the oligos (synthesis or labeling). Data from different microarrays are aligned on a line of slope equal to one in the log-log plots of Fig. 1, which implies a linear relationship between the intensities. In general, replicates show a strong correlation between median intensities, which is an indication of

a good reproducibility of the results. We included in this median the probes with and without poly(A) spacer. No significant difference was found in the intensities from spots with poly(A) and without poly(A) spacer. From this point on, the median intensity of 15 replicates is always used and simply referred to as the intensity of a probe, and because of the good reproducibility we will only discuss the data produced by hybridizations with oligo synthesis a (see Table 3) .

Data analysis with ΔG_{sol}

Next, we consider the relation between the intensities and the corresponding ΔG_{sol} for hybridizations in solution with one or two mismatches. In the case of two mismatches ΔG_{sol} was calculated as the sum of nearest neighbor parameters for individual mismatches, assuming that the presence of two mismatches does not involve additional terms in the free energy, i.e. they do not interact. In the experiment the minimal distance between two mismatches is 5 nucleotides, which is considered sufficient, in first approximation to support the non-interaction assumption. In the calculation of ΔG from the tabulated values of ΔH and ΔS the temperature was set to the experimental value $T = 65°C$.

Figure 2 (a) shows plots of the intensities vs. $\Delta\Delta G_{\text{sol}}$ as taken from the nearest neighbor model with the existing tabulated values for hybridization in solution (see Ref. [6] and references therein). $\Delta\Delta G_{\text{sol}}$ is obtained by subtracting from all free energies that of the PM sequence, which is taken as a reference. As a consequence, for the PM intensities $\Delta\Delta G_{\text{sol}} = 0$. Each plot in Fig. 2 contains 1006 data points obtained from the median value of the 15 replicated spots on each array.

As it is well-known from several studies of melting/hybridization in aqueous solution (see e.g. [7]), the hybridization free energy ΔG_{sol} depends on the buffer conditions, and in particular of the ionic strenght of the solution. Particularly studied was the effect of salt concentration

(NaCl), which is usually assumed to be independent of sequence, but to be dependent on oligonucleotide length. Melting experiments in solution are consistent with the following dependence on Na ions concentrations [7]

$$\Delta G_{\rm sol} = \Delta G_{\rm sol}(1M[Na^+]) - aN \ln[Na^+] \tag{13}$$

where $\Delta G_{\rm sol}(1M[Na^+])$ is measured at 1M NaCl, N is the number of phosphates in the sequence and a a constant. To our knowledge, the salt effect on sequences with internal mismatches has not been investigated yet, as measurements were done at 1M NaCl [6]. However, salt has mostly an effect on interactions with the negatively charged phosphate molecules. It is hence plausible to expect the same type of correction as Eq. (13) also for sequences carrying mismatches. If that is the case, the salt dependence cancels out from $\Delta\Delta G_{\rm sol}$, which is the quantity we are interested in. In the rest of the paper, we will set the value at 1M NaCl in $\Delta G_{\rm sol}$.

Figure 2(a) shows the data for Experiment 1 at three different concentrations, from bottom to top of 50, 500 and 5000 pM. When plotted as functions of $\Delta\Delta G_{\rm sol}$ the data points tend to cluster along single monotonic curves. This already suggests a fair degree of correlation between $\Delta G_{\rm sol}$ and $\Delta G_{\mu\rm array}$. The experiment at 5000 pM shows a pronounced saturation of the intensities, as expected from the Langmuir model (Eq. 2). Sufficiently far from saturation one expects a linear relationship between the logarithm of the intensity and ΔG, as given by Eq. (4). Figure 2 shows that the low concentration data at low intensities follow approximately a straight with the slope $1/RT$ expected from equilibrium thermodynamics at $T = 65°C$, which is the experimental temperature.

However, the global behavior of the three concentrations is at odds with the Langmuir model, which predicts that Intensity vs. free energy plots for different concentrations should saturate at a common intensity value A, as indicated in Fig. 2(b). Although one may expect

some variations on A from experiment to experiment, the data of Figure 2(a) are hard to reconcile with the Langmuir model. We conclude that the hybridization data deviate from the full Langmuir model of Eq. (2), but they are in rather good agreement with its limiting low intensities behavior Eq. (4). In order to obtain estimates of the free energies $\Delta\Delta G_{\mu\text{arrays}}$ from microarray data we will use then Eq. (4) and restrict ourselves to the lower concentration data. The analysis of the higher concentration regime is presented in the supplementary online material.

Fitting the free energy parameters

To fit the 58 parameters of the nearest neighbor model we use the lowest concentration data, i.e. 50 pM. Hereto we applied the algebraic procedure explained in Materials and Methods, which fits the logarithm of the ratios I/I_{PM} and which assumes that the data can be described by Eq. (4). For low concentrations this assumption is expected to be correct for the lower intensities but not for the highest intensities, which deviate from the Langmuir isotherm as shown in Fig. 2. This poses a problem for the fitting procedure since it was designed with the perfect match intensity I_{PM} as a reference (Eq. (7)). One may think to circumvent this problem by restricing the fit to low intensities, for instance only to probes with two mismatches and rewrite Eq. (7) using as reference not I_{PM}, but for instance one of the intensities of a probe with two internal mismatches. This procedure turns out to be of little practical use for our purposes which is to estimate the free energy difference between perfect matching sequences and sequences with one or multiple mismatches and for which the PM reference value is necessary (a more detailed discussion is in the online material).

From the analysis of plots of Intensity vs. $\Delta\Delta G_{\text{sol}}$ (Figure 2) one finds that the PM intensity is systematically lower than that predicted by (Eq. (4)), which is the straight line in Figure 2(a). Hence, the relative intensities I/I_{PM} of the probes that contain mismatches are systematically higher than those predicted by Eq. (4). Consequently, a direct fit of the

experimental data to Eq. (7) underestimates the effect of a mismatch, which will result in free energy penalties that are too small. The result of the fit is shown in Figure 3. One can notice that the $\Delta\Delta G$ range is indeed smaller than the one from hybridization in solution (Figure 2). Moreover, the underestimation of $\Delta\Delta G$ is more severe for probes with two mismatches than for those with only one, since $\Delta\Delta G$ is a sum of contributions per mismatch. This produces a discontinuity of the curve from double to single mismatches. The appearance of this discontinuity is another evidence of the fact that Eq. (4) is not valid in the full range of intensities.

In order to solve this problem, one would need to fit the data with a more general model $I(c, \Delta G)$ that incorporates the observed deviations from Eq. (4). As mentioned above, and as shown explicitely from the data analysis in the supplementary online material, the deviations cannot be described within the general Langmuir model (Eq. (2)). At present, it is not yet clear which alternative model to use for $I(c, \Delta G)$. Moreover, the choice of this model may considerably influence the fitted nearest neighbor parameters. A safer compromize is to start from the observation that Eq. (4) is followed by the large majority of the low concentration data points in Fig. 2. Hence a fit to the low concentration limit of the Langmuir model seems reasonable. Unfortunately, one of the points deviating from Eq. (4) is the PM intensity, which is used as reference measure. In order to calibrate the fit correctly one should reweight the reference PM intensity. We therefore fit the data against Eq. (7) using instead of the actual PM intensity as a reference, a rescaled value $I^*_{PM} = \alpha I_{PM}$, which is the value the PM intensity would have if the data would agree with Eq. (4) in the whole intensity range. We estimate α from the crossing of the 50pM fitting line in Fig. 2(a) with the $\Delta\Delta G = 0$ axis. This estimate is $\alpha = 30$. The effects of a change in α on the fitting parameters will be discussed below.

Figures 4(a-d) show the result of the fit to Eq. (7), using $\alpha = 30$. In the main frames each experiment is fitted independently. In the insets the free energy parameters are obtained

from a simultaneous fit of all 50pM experiments. The latter data produce more accurate parameters, as they come from using 4 independent experiments (the 4 experiments at 50 pM, oligo synthesis a, in Table 3), hence the 58 parameters are obtained on sampling over 1006×4 data points. Both the free energy range and the continuity of the curves in Fig. 4 are now as expected. The data show very little spreading in comparison with the curves in Figure 2(a). A quantification of the spreading for a monotonic curve can be assessed by the Spearman's rank correlation coefficient, which for all four experiments is very close to 1. This is an indicator of the reliability of the the nearest neighbor fitted parameters. The ratio of data points over tuning parameters is large 4024/58, which ought to yield a reliable fit. Moreover, although the data are fitted to a linear model, all four experiments show a clear deviation for the highest intensities. This is an indication against overfitting, which would result in a fully linear curve with erroneous fitting parameters. Therefore we conclude that the deviations from the Langmuir isotherm observed in all four experiments is a robust feature of the system and that the resulting free energy parameters are physically meaningful. We also verified that the free energy parameters obtained from the fit are quite stable whether one fits the whole set of experimental data, or whether the fit is restricted to the lowest intensity scales (e.g. $I/I^*_{PM} \leq 5 \cdot 10^{-3}$) where all data clearly follow Eq. (4). This is because the large majority of experimental points in Fig. 4 are located in the lowest intensity scales, anyhow. Hence, this additional data filtering has little effects on the parameters.

Table 4 shows the free energy parameters $\Delta\Delta G_{\mu\mathrm{array}}$ as obtained from the above fitting procedure. Because of the degeneracies mentioned above (see e.g. Eq. (12) and Ref. [18]) the dinucleotide parameters are not uniquely determined. Triplet parameters are however unique, and these are given in the Table. The parameters are the $\Delta\Delta G$ defined, for instance, as:

$$\Delta\Delta G \begin{pmatrix} \mathrm{A\underline{C}G} \\ \mathrm{T\underline{T}C} \end{pmatrix} = \Delta G \begin{pmatrix} \mathrm{A\underline{C}G} \\ \mathrm{T\underline{T}C} \end{pmatrix} - \Delta G \begin{pmatrix} \mathrm{AAG} \\ \mathrm{TTC} \end{pmatrix} \tag{14}$$

where the upper strand is 5'-3' oriented. The lower strand is the invariant target sequence, the upper strand are the probe sequences. Hence the $\Delta\Delta G$ parameters are measured subtracting the reference perfect match probe. Note that because of this subtraction one has

$$\Delta\Delta G \begin{pmatrix} \mathrm{A\underline{C}G} \\ \mathrm{T\underline{T}C} \end{pmatrix} \neq \Delta\Delta G \begin{pmatrix} \mathrm{C\underline{T}T} \\ \mathrm{G\underline{C}A} \end{pmatrix} \tag{15}$$

as the reference PM sequence is different in the two cases.

Using standard linear regression tools we estimated the error bar on the parameters of Table 4 to be equal to 0.2. In order to compare with existing published data [6] we present in Table 5 the $\Delta\Delta G_{sol}$ for triplets following the same notation as in Table 4. As mentioned before the data in solution are at $T = 65°C$ and 1M [Na$^+$]. Figure 5 shows a plot of the two free energies $\Delta\Delta G_{\mu\mathrm{array}}$ vs. $\Delta\Delta G_{\mathrm{sol}}$. A clear quantitative correlation between the two is observed. The Pearson correlation coefficient is 0.839. In comparing the two sets we note that the 16 mismatches of CC appear to be the most deviating in the two cases.

As discussed above, the fit was done with a rescaled PM intensity, using a factor $\alpha = 30$. We have repeated the analysis for other values of α. Varying α causes a global shift of the data in Table 4 by an α dependent constant. This shift does not affect the slope or correlation of the data in Fig. 5. By using $\alpha = 50$ we found a positive shift of 0.17, while setting $\alpha = 20$ produces a shift of -0.14. These two values of α are our estimate of the largest range of variability for this parameters. In general the procedure of reweighting the PM intensity with α introduces a global error ± 0.2 affecting all parameters in Table 4.

Concluding remarks

During the past decades a considerable amount of research was devoted to the quantification of interactions among hybridizing nucleic acid strands in aqueous solution. This lead to a parametrization, via the nearest-neighbor model, of the contribution to the total free energy

in terms of dinucleotide pairs for perfect matching DNA/DNA [7], RNA/RNA [19] and DNA/RNA [20] duplexes, but also for strands with an internal mismatch [6]. This large amount of data is currently used in various applications as for instance for calculation of DNA melting temperatures or for RNA secondary structure predictions. As it has been widely recognized [6,15,16] a similar effort for quantifying interactions in DNA microarrays is very important. This effort will lead to a better understanding of molecular interactions in DNA microarrays and ultimately on their functioning.

A precise quantification of interactions brings some challenges. First of all many different microarray platforms exist, they differ by the length of probe sequences and the way these are covalently linked to the solid surface. It is not unlikely that interactions between hybridizing strands are of slightly different nature in these different platforms. Hence, one should be careful for instance to generalize the results of this work to, say, Affymetrix GeneChips. In addition, in order to measure accurately interaction parameters, one needs a careful experimental setup in which possible competing reactions, as hybridization between partially complementary strands in solution, are absent. In the case of the present work this was achieved by choosing a single sequence in solution hybridizing with perfect matching probe sequences with one or two internal mismatches. It is difficult to directly fit the free energy parameters from complex biological experiments where the hybridizing solution contain a large number of interacting sequences. This may explain why in some cases poor correlations between ΔG_{sol} and $\Delta G_{\mu\mathrm{array}}$ was reported [10,11]. One of the advantages of the experimental setup chosen in this work is that one can obtain in principle all parameters in a single experiment, as all hybridization reactions with one or two mismatches occur in "parallel" on a single array. However, a drawback is that in this setup one can determine only the free energy and not the contribution of enthalpy and entropy separately, which would allow to extend the parameters to other temperatures. It would be certainly interesting to extend the analysis to other platforms and hybridization conditions.

In the present work we focused on the determination of $\Delta\Delta G$ which is the free energy difference between a perfect matching hybridization and an hybridization where the probe sequences has one or more internal mismatches. Quantifying the effect of internal mismatches is important for a better understanding of cross-hybridization effect, which is the unintended binding of non-perfectly complementary sequences to a given probe. Moreover, this understanding could have some practical consequences for optimal probe design. An advantage of the parameter $\Delta\Delta G$ is that it is insensitive to the free energy initiation parameter (Eq. (5)) and the scaling factor A (Eq. (2), Eq. (4)) and that it is expected to be less sensitive to buffer conditions as ionic salt etc ... The determination of the perfect match parameters ΔG is also possible in principle from microarray experiment but it requires sampling perfect match hybridizations from a large number of target sequences. This requires a different and more complex experimental design.

The present work on custom Agilent arrays shows that there is a strong correlation, also on the quantitative scale, between $\Delta\Delta G_{\text{sol}}$ and $\Delta\Delta G_{\mu\text{array}}$. This correlation is shown in Fig. 5 with explicit free energy values given in Tables 4 and 5. A fit of the interaction parameters from microarray data shows a much better agreement of the data with the thermodynamic models (compare Fig. 2 with Fig. 3). However, in absence of dedicated experiments for the determination of interaction free energies on a DNA microarray, the results of this work suggest that one could use as approximations for them the corresponding hybridization free energies in solution. Recent work [9,15] has addressed the issue of the correlation between ΔG_{sol} and $\Delta G_{\mu\text{array}}$. Ref. [9] considered oligonucleotide microarrays on Codelink activated slides carrying one, two or three mismatches. The data plotted as a function of ΔG_{sol} showed a good agreement with the Langmuir model, implying a fair correlation between ΔG_{sol} and $\Delta G_{\mu\text{array}}$. However the number of data points was insufficient to perform a direct fit of the thermodynamic parameters from the microarray data. Interestingly, the lowest concentration data in Ref. [9] seem to indicate the existence of deviations from the Langmuir model similar

to those observed in Fig. 2(a). Fish et al. [15] performed a series of experiments on oligo sequences in solutions hybridizing to perfect match and to sequence carrying one to multiple mismatches. Their analysis included tandem mismatches, i.e. mismatches on neighboring sequence sites (in our case the minimal distance between mismatches is five nucleotides). An overall correlation between ΔG_{sol} and microarray intensities was observed, implying a correlation between ΔG_{sol} and $\Delta G_{\mu\mathrm{array}}$. In these experiments ΔG_{sol} was measured directly from experiments in solution and did not rely on the nearest-neighbor model parameters. As a correlation between ΔG_{sol} and $\Delta G_{\mu\mathrm{array}}$ has by now been observed in several different microarray platforms, it is fair to expect that such a correlation is a general feature of microarrays. However, an accurate determination of nearest-neighbor parameters in other platforms would be very useful for a better quantification of this correlation.

An interesting issue is the deviation from the low concentration limit of the Langmuir model (Eq. (4)). These deviations cannot be explained by the full model of Eq. (2). There are several underlying approximations in the Langmuir model, as for instance hybridization is always considered two state. The model also assumes that hybridizing strands, apart from forming a duplex, do not further interact with other strands at the surface. Moreover, Eqs. (2) and (4) apply to a system in thermal equilibrium. More investigations are necessary for a better understanding on the deviation from the Langmuir model found in this study. These will involve further experiments in different external conditions, e.g. different temperatures or salt concentrations as well as theoretical analysis, which are left for some future work.

It is interesting to remark that the deviation from the Langmuir model "enhances" the cross-hybridization problem because there is a smaller effect on intensity for a given free energy penalty (smaller slope in Figure 4). As an example, a mismatch with $\Delta\Delta G = 2.5$ kcal/mol (a typical value from Table 4) corresponds to a I/I_{PM} ratio of ≈ 0.02 in the regime governed by the Langmuir model, compared to ≈ 0.2 in the deviating regime. This implies that in the deviating regime a significant fraction of the amount of target binding to a PM probe binds

to a probe carrying one internal mismatch.

Although the origin of these deviations are not known it is remarkable that the data appear to follow approximately two straight lines separated by a sharp kink (Fig. 4). Although extensions of the Langmuir model in the context of DNA microarrays have been discussed (see e.g. [21]) we are unaware of isotherms which could have a shape as shown in Fig. 4. The presence of a second straigth line in the log plot implies that in this range the data still follow the thermodynamic model of Eq. (4) but with a different "effective" temperature than the experimental one. A linear regression to the data yields $T_{\rm eff} \approx 850K$, which is higher than the experimental temperature. It is interesting to point out that recent analysis [12,13] of Affymetrix GeneChip data use Langmuir model with $\Delta G_{\rm sol}$ rescaled to higher effective higher temperatures. A better understanding of the regime governed by an effective temperature may provide new insights on this issue.

Acknowledgements

We thank Kizi Coeck and Karen Hollanders for help with the experiments. We aknowledge financial support from the Fonds voor Wetenschappelijk Onderzoek (FWO) under grant G.03111.08.

References

[1] Brown, P. O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature* **21**, 33.

[2] Stoughton, R. B. (2005) Applications of DNA microarrays in biology. *Annu Rev Biochem* **74**, 53–82.

[3] Bloomfield, V. A., Crothers, D. M. and Tinoco, Jr., I. (2000) Nucleic Acids Structures,

Properties and Functions, University Science Books, Mill Valley, .

[4] Borer, P. N., Dengler, B., Tinoco, I. and Uhlenbeck, O. C. (1974) Stability of ribonucleic acid double-stranded helices. *J Mol Biol* **86**, 843–853.

[5] Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. and Turner, D. H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci.* **83**, 9373–9377.

[6] SantaLucia, Jr., J. and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 415.

[7] SantaLucia Jr., J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.* **95**, 1460.

[8] Fotin, A. V., Drobyshev, A. L., Proudnikov, D. Y., Perov, A. N. and Mirzabekov, A. D. (1998) Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucleic Acids Res* **26**, 1515–1521.

[9] Weckx, S., Carlon, E., De Vuyst, L. and Van Hummelen, P. (2007) Thermodynamic Behavior of Short Oligonucleotides in Microarray Hybridizations Can Be Described Using Gibbs Free Energy in a Nearest-Neighbor Model. *J. Phys. Chem. B* **111**, 13583.

[10] Naef, F. and Magnasco, M. O. (2003) Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E* **68**, 011906.

[11] Zhang, L., Miles, M. F. and Aldape, K. D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotech.* **21**, 818.

[12] Held, G. A., Grinstein, G. and Tu, Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl. Acad. Sci.* **100**, 7575.

[13] Carlon, E. and Heim, T. (2006) Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays. *Physica A* **362**, 433.

[14] Fiche, J. B., Buhot, A., Calemczuk, R. and Livache, T. (2007) Temperature effects on DNA chip experiments from surface plasmon resonance imaging: Isotherms and melting curves. *Biophys. J.* **92**, 935.

[15] Fish, D. J., Horne, M. T., Brewood, G. P., Goodarzi, J. P., Alemayehu, S., Bhandiwad, A., Searles, R. P. and Benight, A. S. (2007) DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. *Nucleic Acids Res* **35**, 7197–7208.

[16] Pozhitkov, A. E., Tautz, D. and Noble, P. A. (2007) Oligonucleotide microarrays: widely applied–poorly understood. *Brief Funct Genomic Proteomic* **6(2)**, 141–148.

[17] Peyret, N., Seneviratne, P. A., Allawi, H. T. and SantaLucia Jr., J. (1999) Nearest-neighbor Thermodynamics and NMR of DNA sequences with internal AA, CC, GG and TT mismatches. *Biochemistry* **38**, 3468.

[18] Gray, D. M. (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. I. Simple sets of independent sequences and the influence of absent nearest neighbors. *Biopolymers* **42**, 783–793.

[19] Xia, T., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C. and Turner, D. H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735.

[20] Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M. and Sasaki, M. (1995) Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry* **34**, 11211.

[21] Vainrub, A. and Pettitt, B. M. (2002) Coulomb blockage of hybridization in two-dimensional DNA arrays. *Phys Rev E* **66**, 041905.

Table 4. Free energy differences $\Delta\Delta G$ unique parameters obtained from fitting microarray data to Eq. (7). The data refer to triplets with central mismatching nucleotides and flanking matching nucleotides. The convention is that the numbers correspond for say, a mismatch $\frac{AGT}{TTA}$ to a free energy difference $\Delta G\left(\frac{AGT}{TTA}\right) - \Delta G\left(\frac{AAT}{TTA}\right)$. The upper strand has orientation 5' - 3'. The error bar on the parameters is 0.2.

X \ Y		A	C	G	T		A	C	G	T		A	C	G	T		A	C	G	T
A	$\frac{XAY}{X'AY'}$	2.2	2.0	2.4	2.2	$\frac{XAY}{X'CY'}$	3.0	2.8	3.0	3.0	$\frac{XAY}{X'GY'}$	2.5	1.8	2.5	2.2	$\frac{XCY}{X'AY'}$	2.4	2.2	2.4	2.5
C		2.3	2.1	2.5	2.4		3.0	2.8	3.0	3.0		2.5	1.7	2.5	2.1		2.4	2.2	2.4	2.5
G		1.9	1.8	2.2	2.0		2.7	2.5	2.7	2.7		2.4	1.6	2.4	2.0		2.0	1.8	2.1	2.1
T		2.2	2.1	2.5	2.3		3.1	2.9	3.1	3.1		2.4	1.7	2.5	2.1		2.4	2.2	2.4	2.5
A	$\frac{XCY}{X'CY'}$	3.9	3.4	3.4	4.0	$\frac{XCY}{X'TY'}$	2.5	2.4	2.4	2.8	$\frac{XGY}{X'AY'}$	1.5	1.3	1.7	1.7	$\frac{XGY}{X'GY'}$	2.4	1.8	2.3	1.9
C		3.4	3.0	2.9	3.5		2.4	2.3	2.3	2.7		1.7	1.6	1.9	1.9		2.7	2.1	2.6	2.2
G		3.1	2.7	2.7	3.2		2.5	2.5	2.5	2.8		1.1	0.9	1.3	1.3		2.5	1.9	2.4	2.0
T		3.8	3.4	3.3	3.9		2.5	2.5	2.4	2.8		1.7	1.6	2.0	2.0		2.8	2.2	2.7	2.3
A	$\frac{XGY}{X'TY'}$	2.0	1.8	1.9	1.9	$\frac{XTY}{X'CY'}$	3.5	3.6	3.1	3.2	$\frac{XTY}{X'GY'}$	2.2	2.2	2.0	2.4	$\frac{XTY}{X'TY'}$	2.3	2.4	2.0	2.2
C		1.6	1.4	1.5	1.5		3.2	3.3	2.8	3.0		2.3	2.3	2.1	2.5		2.1	2.2	1.7	2.0
G		1.8	1.7	1.8	1.7		3.1	3.2	2.8	2.9		2.4	2.4	2.2	2.6		2.4	2.5	2.1	2.4
T		1.6	1.4	1.6	1.5		3.2	3.3	2.9	3.0		2.3	2.3	2.1	2.5		2.2	2.3	1.9	2.1

Table 5. Data as in Table 4 using the nearest neighbor parameters obtained from melting experiments in solution (see [6] and references therein). The data are at $T = 65°$ C and at 1 M [Na$^+$].

X \ Y		A	C	G	T		A	C	G	T		A	C	G	T		A	C	G	T
A	$\frac{XAY}{X'AY'}$	1.3	2.0	2.3	2.0	$\frac{XAY}{X'CY'}$	2.9	3.6	3.5	2.6	$\frac{XAY}{X'GY'}$	2.3	1.7	2.9	1.8	$\frac{XCY}{X'AY'}$	1.4	1.8	2.1	1.8
C		1.6	2.3	2.6	2.2		3.5	4.2	4.1	3.2		2.7	2.0	3.2	2.1		2.2	2.6	3.0	2.6
G		1.6	2.3	2.6	2.3		3.1	3.8	3.7	2.9		2.6	2.0	3.2	2.1		2.1	2.5	2.9	2.6
T		1.1	1.8	2.1	1.8		3.0	3.7	3.6	2.7		2.5	1.8	3.0	1.9		1.9	2.3	2.6	2.3
A	$\frac{XCY}{X'CY'}$	3.4	4.3	4.4	4.5	$\frac{XCY}{X'TY'}$	2.2	2.5	2.3	2.2	$\frac{XGY}{X'AY'}$	0.8	0.8	1.3	1.0	$\frac{XGY}{X'GY'}$	2.1	1.6	2.6	1.7
C		3.6	4.5	4.5	4.7		2.7	3.0	2.8	2.7		1.6	1.6	2.1	1.7		2.8	2.2	3.2	2.4
G		3.1	4.0	4.1	4.2		2.3	2.6	2.4	2.4		0.7	0.7	1.1	0.8		2.1	1.5	2.6	1.7
T		2.6	3.5	3.6	3.7		2.1	2.4	2.2	2.2		1.4	1.3	1.8	1.5		2.3	1.7	2.8	1.9
A	$\frac{XGY}{X'TY'}$	2.0	1.7	1.7	1.7	$\frac{XTY}{X'CY'}$	3.3	3.6	3.6	3.1	$\frac{XTY}{X'GY'}$	2.4	2.2	2.4	2.5	$\frac{XTY}{X'TY'}$	2.5	2.8	2.4	2.6
C		1.6	1.3	1.3	1.3		3.6	4.0	4.0	3.4		2.5	2.2	2.4	2.6		2.3	2.6	2.2	2.4
G		1.4	1.1	1.1	1.1		3.6	3.9	3.9	3.4		2.6	2.4	2.6	2.7		2.4	2.8	2.4	2.6
T		1.4	1.1	1.1	1.1		3.2	3.6	3.6	3.0		2.5	2.2	2.4	2.6		1.8	2.2	1.7	2.0

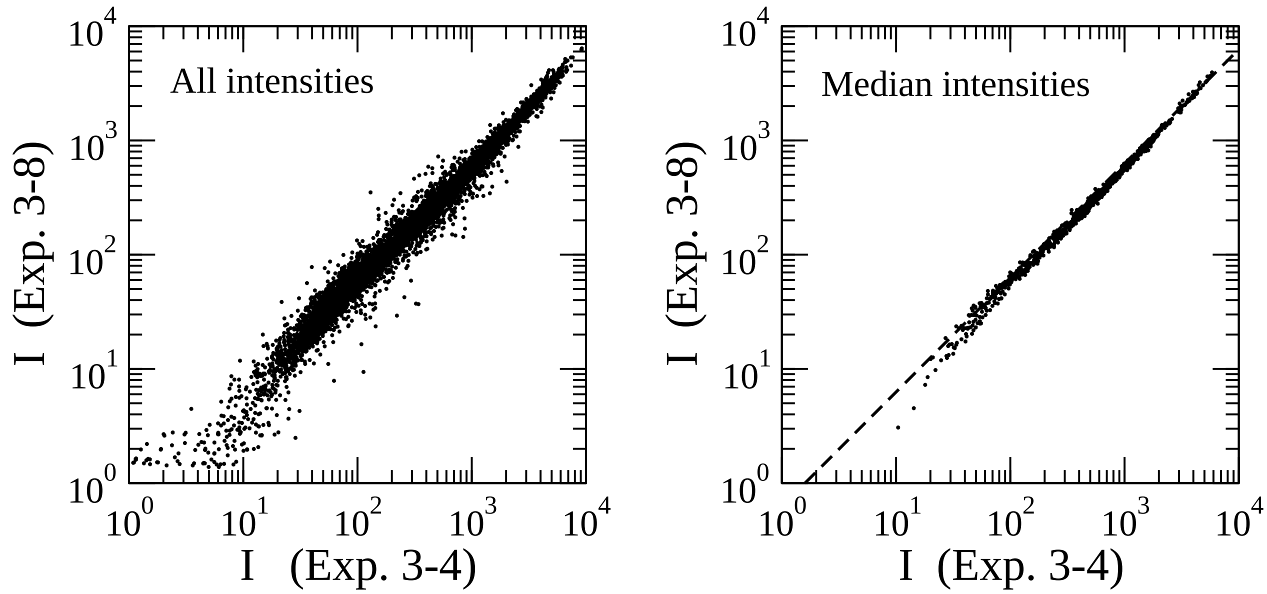

FIGURE 1

Fig. 1. Correlation plots for intensities in two replicated experiments at 50 pM for oligo 3a (x-axis) and oligo 3b (y-axis); these are the experiments 3-4 and 3-8 in Table 2. The left plot shows the total intensities and the right plot concerns only the median intensities taken for the 15 replicated spots. The dashed line has slope equal to one.

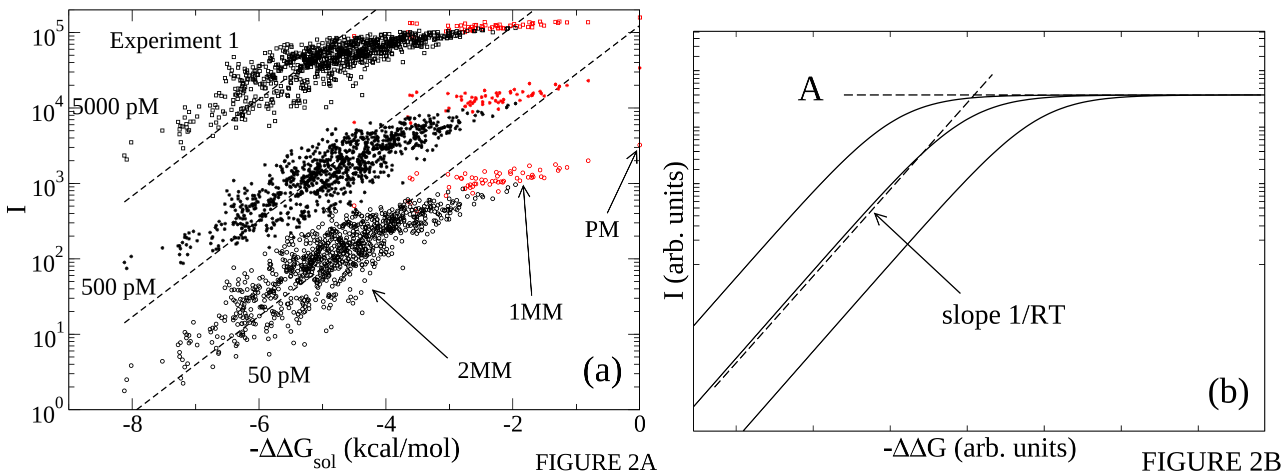

Fig. 2. (a) Plot of the intensities as functions of $\Delta\Delta G_{\mathrm{sol}}$, the difference of hybridization free energy with respect to the perfect match free energies, from nearest neighbor free energies obtained from melting experiments in solution. With this choice of parameters the perfect match is **located at** $\Delta\Delta G = 0$**.** The different plots correspond to concentrations of 50, 500 and 5000 pM (from bottom to top). The lines drawn have slopes corresponding to $1/RT$, with $T = 65°C = 338K$ the experimental temperature. (b) Behavior of three concentration data as predicted from the Langmuir model (Eq. (2)).

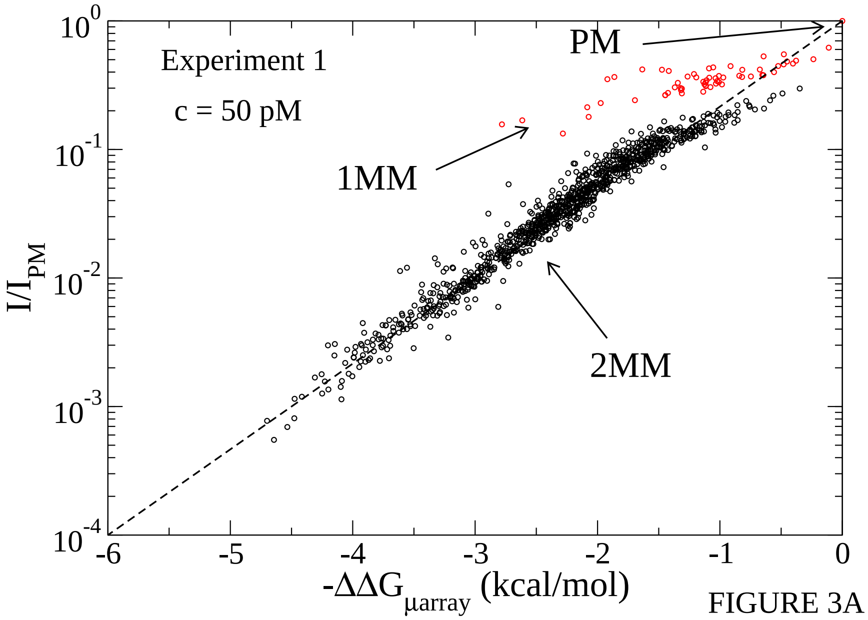

Fig. 3. Ratios of Intensities and perfect match intensities vs. $\Delta\Delta G_{\mu\text{array}}$, the **relative** hybridization free energy between two strands as obtained from a fit to Eq. (7). Three distinct groups of points are indicated: PM for perfect match, 1MM for probes with a single internal mismatch and 2MM for probes with two mismatches. The dashed line in is drawn as a guide to the eye.

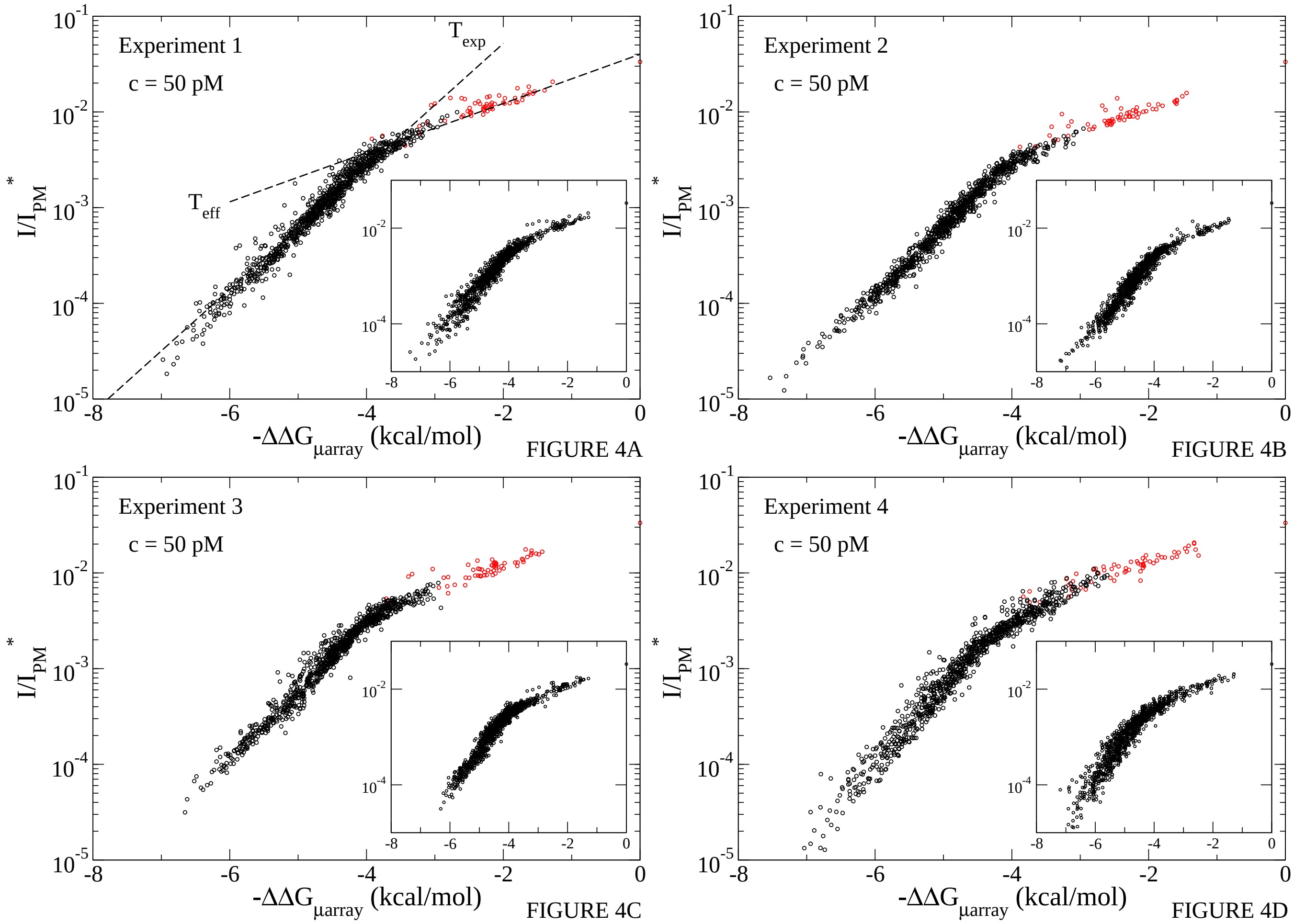

Fig. 4. Plot of $I/I^*_{\rm PM}$ where $I^*_{\rm PM} = \alpha I_{\rm PM}$ (where we took $\alpha = 30$ as explained in the text) as function of the nearest neighbor fitted $\Delta\Delta G_{\mu\rm marray}$. The alignment of the intensities onto single monotonic curves is a proof of the good quality of the fits. In the main frame the four different experiments where fitted separately. The insets show the date from intensities of each experiments, but the fit was done globally on all experiments at 50 pM. As a measurement of the goodness of the fits the Spearman's rank correlation coefficient was used. This coefficient is for the main frames plots (a-d): 0.9860 0.9911 0.9866 and 0.9867. For the four plots in the insets the correlation coefficients are: 0.9732 0.9705 0.9748 and 0.9699. The two straight lines in the first main frame correspond to slopes $1/RT$ where we took $T_{exp} = 65^\circ$ C $= 338K$ for the experimental temperature and $T_{eff} = 850$K.

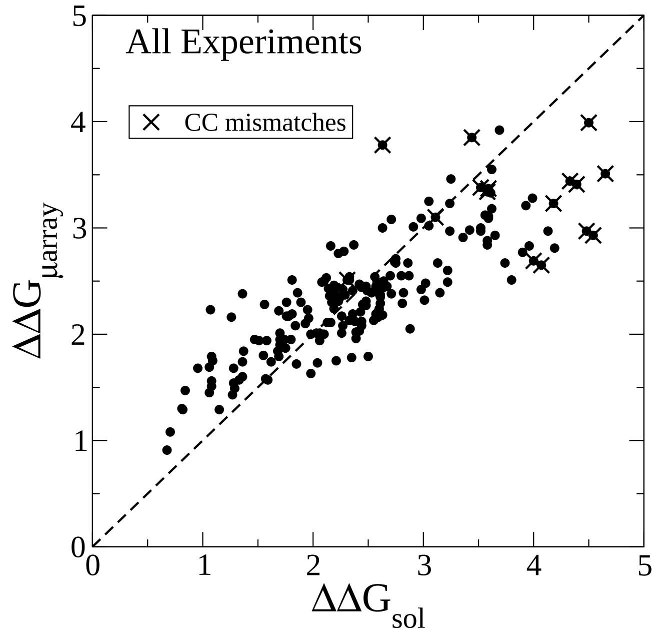

FIGURE 5

Fig. 5. Comparison of data in Table 4 and 5: the free energy differences between a perfect matching hybridization and hybridization with an internal mismatch as obtained from data from Ref. [6] ($\Delta\Delta G_{\mathrm{sol}}$) and from a fit of the microarray data ($\Delta\Delta G_{\mu\mathrm{array}}$). The results show a good quantitative correlation between the two quantities: the Pearson correlation coefficient is 0.839